Analyse & Kritik 27/2005 (© Lucius & Lucius, Stuttgart) 106–112

Jason McKenzie Alexander

The Evolutionary Foundations of Strong Reciprocity

Abstract: Strong reciprocators possess two behavioural dispositions: they are willing to bestow benefits on those who have bestowed benefits, and they are willing to punish those who fail to bestow benefits according to some social norm. There is no doubt that peoples' behaviour, in many cases, agrees with what we would expect if people are strong reciprocators, and Fehr and Henrich argue that many people are, in fact, strong reciprocators. They also suggest that strongly reciprocal behaviour may be brought about by specialised cognitive architecture produced by evolution. I argue that specialised cognitive architecture can play a role in the production of strongly reciprocal behaviour only in a very attenuated sense, and that the evolutionary foundations of strong reciprocity are more likely cultural than biological.

To say that people cooperate, trust, offer more than is logically or strategically necessary, and punish those who attempt to short-change them, even when incuring a cost to themselves, is not to say anything new. What *is* new are current approaches to explaining such behaviour, where "explain" here means identifying the ultimate causal factors responsible for it. The task is to identify the ultimate, not proximate, causal factors, because it is entirely straightforward to identify some proximate causes: the beliefs and desires of the agent at the time of acting. Why did so-and-so choose to offer more than is logically or strategically necessary? Because (one might say) that person believes that that is what they ought to do, and there is no overriding reason to not do what they ought to do.[1] Dissatisfaction with this type of explanation arises because it leaves unanswered the question of *where* those beliefs ultimately came from. Why those beliefs and not others? To say they were learned seems inadequate because it simply pushes back the inquiry a stage: if the beliefs were learned, from whom were they learned, and why did *that* person hold those beliefs and choose to propagate them by teaching?

A search for the ultimate causal factors is thus a search for the real reasons that people cooperate, trust, and so on. An evolutionary psychological account is a natural one to propose, since it locates the source of certain key motivating beliefs, desires, and dispositions in the fitness advantages conferred. Along this

[1] Setting aside obvious complexities introduced by *akrasia*.

line, Fehr and Henrich (2003) consider the evolutionary foundations of strong reciprocity.

To say that an individual is a strong reciprocator is to say that person possesses two behavioral dispositions. First, that person is willing "to sacrifice resources to bestow benefits on those who have bestowed benefits" and, second, that person is willing to "sacrifice resources to punish those who are not bestowing benefits in accordance with some social norm" (Fehr/Henrich 2003, 57). It is a brute fact that individual behaviour in economic experiments and other contexts is, in many cases, equivalent to that which would be produced if individuals possessed the above dispositions and acted accordingly. Given the importance of social norms and reciprocal behaviour for explaining and understanding human social behaviour, key questions are then (a) is the observed behavior of individuals in these contexts actually generated by such dispositions? (b) if so, how are these dispositions acquired? and (c) is the behaviour produced by these dispositions adaptive? Fehr and Henrich take the experimental record to have answered the first question in the affirmative, with the existence of strong reciprocity among humans having clearly been documented in recent years.[2] And given this, the debate over strong reciprocity has thus turned to questions of whether it is, ultimately, adaptive or maladaptive, and what evolutionary forces have served to make humans strong reciprocators. In what follows, I will concentrate on the question of how the dispositions underlying strong reciprocity are acquired.

One should note that strong reciprocity, as defined above, is officially neutral with respect to its ultimate cause. The necessary dispositions for strongly reciprocal behaviour may have been acquired via purely cultural means, through learning or indoctrination, or may have arisen as a product of habit. Alternatively, the dispositions for strong reciprocity may be embedded in one or more evolved psychological mechanisms.[3] Of course, these two options are not mutually exclusive, as the products of cultural conditioning may work hand-in-hand (or at cross-purposes) with evolved psychological mechanisms to cause (or inhibit) strongly reciprocal behaviour.

The fact that strongly reciprocal behaviours are capable of being produced by either evolved psychological mechanisms or purely cultural productions is im-

[2] Which, really, ought not be surprising. The fact that people possess such dispositions is obvious and can be discovered through introspection. I suspect the primary reason for needing to demonstrate and document the existence of strong reciprocity, as defined, derives from the theoretical legacy of behaviorism and instrumentalism in positive economics, coupled with the all-too-common view that the self-interested nature of *homo economicus* translates into utility functions which monotonically increase in the amount of resources possessed. There is no reason, however, for thinking that the self-interested nature of *homo economicus* precludes preferences like those attributed to strong reciprocators.

[3] This way of putting the alternatives is, of course, far too crude if one thinks that a great deal of culture occurs as a natural product of evolved psychological mechanisms. In general, I think Tooby and Cosmides (1992) are right in stressing that the 'Standard Social Science Model' overplays the contrast between nature and nurture as competing, incompatible explanations of human behaviour. My point here, though, is that even if we accept the empirically implausible view that one can perfectly differentiate between nature and nurture as alternate sources from which one may acquire dispositions to act, strong reciprocity, as defined, remains perfectly agnostic as to which is responsible for generating the dispositions.

portant to keep in mind when considering the formal evolutionary models that purportedly show how strongly reciprocal behaviour could have evolved. One noteworthy feature of many evolutionary game theoretic models is that the mathematical formalism readily admits both a biological and cultural interpretation. The classic example of this, of course, is the replicator dynamics of Taylor and Jonker (1978) which, although initially derived in a biological context to provide dynamics for Maynard Smith's notion of an evolutionary stable strategy, was later shown to also provide dynamics for some models of cultural evolution via imitative learning. (See Weibull 1995 for one such derivation.) As many other evolutionary models also have this dual interpretive nature, providing a formal model which shows that strongly reciprocal behaviour could have evolved does not necessarily give one reason to think that evolved psychological mechanisms for strong reciprocity exist.

All things considered, though, Fehr and Henrich do seem to favour viewing strongly reciprocal behaviour as at least partly generated by evolved psychological mechanisms. Discussing evidence from contemporary foraging populations and primate studies, they state that "we think that this evidence should lead evolutionarily minded scholars to predict that humans should be equipped with specialized cognitive machinery capable of distinguishing low-frequency interactants ... from long-term repeated interactants" (Fehr/Henrich 2003, 72). And, slightly later in the same paragraph, they further state that "we think that much of the laboratory-observed behaviour results from adaptive processes acting on human psychology over the course of hominid evolution."

However, strongly reciprocal behaviour is only *partly* generated by such mechanisms. Because subjects quickly adjust their helping or punishing behaviour when changes are made to the cost of helping or punishing, Fehr and Henrich are skeptical of the view that "a cognitively inaccessible mechanism drives the baseline pattern of reciprocal responses" (Fehr/Henrich 2003, 68). Rational deliberation and complex calculation of expected costs and benefits are an indispensible part of the story behind strongly reciprocal behaviour.

So if evolved psychological mechanisms are behind strongly reciprocal behaviour, it is only in an attenuated sense. This view meshes well with comments from some of the economic experiments in which strong reciprocity is invoked to explain behaviour. Fehr notes that students who participate in one-shot prisoner's dilemmas are "often disappointed because they failed to exhaust large parts of the potential gains from cooperation" (Fehr/Henrich 2003, 62). The "disappointment" reported indicates that students had an explicit goal they were trying to bring about[4]—in this case, maximisation of individual payoff—and were seeking to establish, by whatever means possible in the limited strategic context of the game, some sort of pattern which would allow that to occur. These reports suggest that the 'disposition' to cooperate in the prisoner's dilemma de-

[4] Disappointment, after all, is the dissatisfaction resulting from unrealized expectations. It would be odd to use "disappointment" to describe the outcome resulting when an evolved psychological mechanism fails to operate as designed or fails to achieve its goal. We do not feel disappointment when our spatial relations module is misled by optical illusions, or when our theory of mind module (Tooby/Cosmides 1992, 113) mistakenly attributes beliefs when we are really just conversing with a machine over the internet.

rives more from rational individuals seeking to maximise their payoffs than from the blind operation of evolved psychological mechanisms for strong reciprocity.

If 'specialized cognitive architecture' plays an attenuated role in the production of strongly reciprocal behaviour, just how much of a role does it play? Consider, first, Fehr and Henrich's definition of strong reciprocity a bit more closely. Strong positive reciprocity, occuring when a person is willing "to sacrifice resources to bestow benefits on those who have bestowed benefits", is a response prompted by suitably interpreted, previously observed behaviour of others. Likewise, strong negative reciprocity, occuring when a person is willing to "sacrifice resources to punish those who are not bestowing benefits in accordance with some social norm", is a response prompted by suitably interpreted, previously observed behaviour of others when the reciprocator also possesses a belief concerning which social norm governs the current type of interaction. In both cases, there is a great deal of work being performed by the qualifier "suitably interpreted". Human behaviour and the objects distributed do not wear the attributes of 'being bestowed' and 'benefit' on their sleeve: the agent has to engage in an explicit act of interpretation to see them a certain way. The reciprocator needs to believe that the things which have been bestowed are *benefits*, and also that the benefits have been *bestowed*, that is, intentionally distributed by the other person as a gift.[5]

Since any object can be converted into a benefit in the presence of the right social convention or context, any specialized cognitive architecture designed for recognition of benefits would have to be very general and flexible indeed. While it is conceivable that we would have special cognitive machinery for handling certain kinds of resource acquisition and distribution problems—because those resources would have naturally and frequently occurred in the environment of evolutionary adaptiveness—the definition of strong reciprocity requires the dispositions to apply for *arbitrary* benefits. It seems, then, unlikely that this part of the production of strongly reciprocal behavior (the classification of objects as benefits) depends on specialized cognitive machinery, except for that specialized cognitive machinery which generally handles beliefs.

A similar point can also be made regarding the judgment that someone is bestowing benefits (for positive strong reciprocity) or failing to bestow benefits (for negative strong reciprocity). What determines whether a person is bestowing benefits, or failing to bestow benefits, is their intentional state, that is, their beliefs and desires. And, hence, there is no specialized cognitive machinery relevant for this part of the production of strongly reciprocal behaviour other than that which handles attribution of beliefs and desires to others.

Might there be specialized cognitive machinery for handling social norms? Possibly. It is certainly the case that a module for handling social norms is less fantastic than some of the other psychological modules which have been postulated to exist. But just what is a social norm? Bicchieri (2005) defines a social norm as follows:

[5] It is important to stress the role played by interpretation and individual belief because, should either of these two belief-states on the agent's behalf fail to obtain, the agent's failure to bestow benefits in turn does not preclude him or her from being a strong reciprocator.

> R is a social norm in a population P if there exists a sufficiently large subset $P_{\text{cf}} \subseteq P$ such that, for each individual $i \in P_{\text{cf}}$:
>
> *Contingency:* i knows that a rule R exists and applies to situations of type S;
>
> *Conditional preference:* i prefers to conform to R in situations of type S on the condition that:
>
> *(a) Empirical expectations:* i believes that a sufficiently large subset of P conforms to R in situations of type S;
>
> and either
>
> *(b) Normative expectations:* i believes that a sufficiently large subset of P expects i to conform to R in situations of type S;
>
> or
>
> *(b') Normative expectations with sanctions:* i believes that a sufficiently large subset of P expects i to conform to R in situations of type S, prefers i to conform, and may sanction behaviour.

A social norm exists exactly when a sufficiently large subset of the population shares a particular set of beliefs and desires. There's no need, then, for there to be specialized cognitive machinery for handling social norms beyond that which generally handles beliefs and desires.

The above definition of "social norm" fits relatively well, I suspect, with what Fehr and Henrich have in mind when they speak of strongly reciprocal behaviour. But I don't think we can use Bicchieri's definition of "social norm" to make sense of Fehr and Henrich's definition of strong negative reciprocity and still identify correctly all instances of strong negative reciprocity. Recall the definition of strong negative reciprocity: a person must be willing "to sacrifice resources to punish those who are not bestowing benefits in accordance with some social norm". Do Fehr and Henrich really mean to require behaviour in accordance with some social norm (call this the *externalist* reading), or behaviour in accordance with the person's belief that some social norm exists (call this the *internalist* reading)?

Given that social norms are capable of being misunderstood or misapplied by individuals, the externalist reading of Fehr and Henrich's requirement means that individuals who seem to engage (or fail to engage) in punitive behavior for good reason are not actually strong reciprocators. For example, if an individual incorrectly believes that a social norm applies (when none do), and engages in punitive behavior, that person is not a strong reciprocator because, as a matter of fact, no social norm requiring that benefits be bestowed in that particular circumstance exists. Likewise, an individual may incorrectly believe that a social norm does *not* apply to the particular interaction (when one really does), and hence choose not to sacrifice resources to punish behaviour, thus also failing to be a strong reciprocator. But both of these seem to be (at least to me) natural instances in which to identify the agent as a strong reciprocator: he is certainly

acting as a strong reciprocator would if all of his beliefs were true. It just happens to be the case that his beliefs in both instances (that a social norm applies and does not apply, respectively) are false. On the internalist reading of Fehr and Henrich's requirement, both instances are compatible with the individual being a strong reciprocator. Since I think what matters is whether individuals have good reason for engaging in punitive behaviour, I favour the internalist reading.

But under the internalist reading, the need to posit special cognitive machinery for handling social norms disappears, since social norms play no real part in making someone a strong reciprocator—it all has to do with whether that person *believes* a social norm exists. Under the internalist reading, there could actually be no social norms at all (because, under Bicchieri's definition, no sufficiently large subset exists) yet everyone could be a strong reciprocator because they all believe (in error) that certain norms do, in fact, exist. Under the internalist reading, the only specialized cognitive machinery required is that which generally handles beliefs.

How, then, are we to make sense of Fehr and Henrich's view that "much of the laboratory-observed behaviour results from adaptive processes acting on human psychology over the course of hominid evolution"? In working through the definition of strong reciprocity, we've seen that, for each of the key terms involved ("benefits", "bestowing", "punishing", and "social norm") the only viable candidates for specialized cognitive architecture is that which is generally involved in handling beliefs, desires, and the attribution of beliefs and desires to others. The attenuated sense in which evolved psychological mechanisms play a role in strong reciprocity, then, seems very attenuated indeed.

Our evolved psychological nature no doubt does play an important part in strong reciprocity. Violations of social norms trigger anger or outrage, and these emotions provide powerful motivations to act, even to act in ways which fail to maximise an actor's expected utility. However, if emotions (or how dispositions to act are stored in the brain) are the primary place where our evolved psychological nature becomes relevant in explaining strongly reciprocal behaviour, it now seems that evolutionary psychology is no longer giving the real reasons for why people cooperate, trust, punish, and so on. The real reasons people cooperate, trust, punish, and so on are found in the beliefs and desires of people, and their attributions of beliefs and desires to others.

In recent years, interest in evolutionary psychology has led many to hypothesize special cognitive machinery for a panoply of human capabilities. People now talk about "a face recognition module, a spatial relations module, a rigid object mechanics model, a tool-use module, a fear module, a social-exchange module, an emotion-perception module, a kin-oriented motivation module, an effort allocation and recalibration module, a child-care module, a social-inference module, a sexual-attraction module, a semantic-inference module, a friendship module, a grammar acquisition module, a communication-pragmatics module, a theory of mind module, and so on" (Tooby/Cosmides 1992, 113). It is, of course, ultimately an empirical question to what extent we carry around specialized cognitive architecture tailor-made to handle the tasks we face. However, given the centrality of beliefs, desires, and the attributions of beliefs and desires, in strong

reciprocity, the *cultural* evolutionary foundations of human altruism are equally, if not more, important.

Bibliography

Bicchieri, C. (2005), *The Grammar of Society: The Nature and Dynamics of Social Norms*, Cambridge

Fehr, E./J. Henrich (2003), Is Strong Reciprocity a Maladaptation? On the Evolutionary Foundations of Human Altruism, in: P. Hammerstein (ed.), *Genetic and Cultural Evolution of Cooperation*, Dahlem Workshop Reprint 90, Cambridge/MA, 55–82

Taylor, P. D./L. B. Jonker (1978), Evolutionary Stable Strategies and Game Dynamics, in: *Mathematical Biosciences 40*, 145–156

Tooby, J./L. Cosmides (1992), The Psychological Foundations of Culture, in: *The Adapted Mind: Evolutionary Psychology and the Generation of Culture*, Oxford, 19–136

Weibull, J. W. (1995), *Evolutionary Game Theory*, Cambridge/MA